Supplementary materials for *Eliminating Bounded Gradient or Variance Assumption for Sign Stochastic Gradient Descents*

A Technical Lemmas

This part collects several necessary technical lemmas.

Lemma 5 *For random variables $\{\zeta_1, \zeta_2, \ldots, \zeta_M\}$ and $\{c_1, c_2, \ldots, c_M\}$, it holds*

$$\mathcal{P}(|\sum_{i=1}^{M} \zeta_i| \geq \sum_{i=1}^{M} c_i) \leq \sum_{i=1}^{M} \mathcal{P}(|\zeta_i| \geq c_i). \tag{15}$$

Lemma 6 ([Robbins and Siegmund, 1971]) *Let $\mathscr{F} = (\mathcal{F}^k)_{k\geq 0}$ be a sequence of sub-sigma algebras of $\mathcal{F}$ such that $\forall k \geq 0$, $\mathcal{F}^k \subset \mathcal{F}^{k+1}$. Define $\ell_+(\mathscr{F})$ as the set of sequences of $[0, +\infty)$-valued random variables $(\xi_k)_{k\geq 0}$, where ξ_k is $\mathcal{F}^k$ measurable, and $\ell_+^1(\mathscr{F}) := \{(\xi_k)_{k\geq 0} \in \ell_+(\mathscr{F}) | \sum_k \xi_k < +\infty \text{ a.s.}\}$. Let $(\alpha_k)_{k\geq 0}, (v_k)_{k\geq 0} \in \ell_+(\mathscr{F})$, and $(\eta_k)_{k\geq 0}, (\xi_k)_{k\geq 0} \in \ell_+^1(\mathscr{F})$ be such that $\mathbb{E}(\alpha_{k+1}|\mathcal{F}^k) + v_k \leq (1+\xi_k)\alpha_k + \eta_k$. Then $(v_k)_{k\geq 0} \in \ell_+^1(\mathscr{F})$ and α_k converges to a $[0, +\infty)$-valued random variable a.s..*

B Proof of Lemma 2

Noticing $1 + x \leq \exp(x)$ for any $x \in \mathbb{R}$, $\xi_{k+1} \leq \exp(\eta_k)\xi_k + \delta_k$. Direct computations then yield the result.

C Proof of Lemma 5

Consider the sets $S_i := \{\omega \mid |\zeta_i(\omega)| \leq c_i\}$ and $S := \{\omega \mid |\sum_{i=1}^M \zeta_i(\omega)| \leq \sum_{i=1}^M c_i\}$. The triangle inequality means

$$S_1 \bigcap S_2 \bigcap \ldots \bigcap S_M \subseteq S.$$

That means

$$S^c \subseteq S_1^c \bigcup S_2^c \bigcup \ldots \bigcup S_M^c.$$

Then we have

$$\mathcal{P}(S^c) \leq \sum_{i=1}^{M} \mathcal{P}(S_i^c).$$

Noticing that $\mathcal{P}(S^c) = \mathcal{P}(|\sum_{i=1}^m \zeta_i| \geq \sum_{i=1}^M c_i)$ and $\mathcal{P}(S_i^c) = \mathcal{P}(|\zeta_i| \geq c_i)$, we then prove the result.

D Proof of Lemma 4

Proof of (10): The Lipschitz gradient of f gives us

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \leq -\gamma_k \langle \nabla f(\mathbf{x}^k), \mathrm{Sign}(\mathbf{v}^k) \rangle + \frac{Ld\gamma_k^2}{2}. \tag{16}$$

Taking conditional expectation on both sides of (16) on χ^k, we then have

$$\mathbb{E}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \mid \chi^k\right) \leq -\gamma_k \|\nabla f(\mathbf{x}^k)\|_1 + \frac{Ld\gamma_k^2}{2} + 2\gamma_k \times \sum_{j=1}^{d} |[\nabla f(\mathbf{x}^k)]_j| \cdot \mathcal{P}\{\mathbf{v}_j^k \neq [\mathrm{Sign}(\nabla f(\mathbf{x}^k))]_j\}. \tag{17}$$

Substituting condition (9) into (17), we then get

$$\mathbb{E}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \mid \chi^k\right) \leq -\gamma_k \|\nabla f(\mathbf{x}^k)\|_1 + \frac{Ld\gamma_k^2}{2} + 2a_1\gamma_k\delta_k f(\mathbf{x}^k) + 2a_2\gamma_k\delta_k. \tag{18}$$

We then derive

$$\mathbb{E}\left(f(\mathbf{x}^{k+1}) \mid \chi^k\right) + \gamma_k \|\nabla f(\mathbf{x}^k)\|_1 \leq (1 + 2a_1\gamma_k\delta_k) f(\mathbf{x}^k) + \frac{Ld\gamma_k^2}{2} + 2a_2\gamma_k\delta_k. \tag{19}$$

Applying Lemma 3 to (19), we then obtain the a.s. convergence result of $(f(\mathbf{x}^k))_{k\geq 0}$. Taking total expectations on both sides of (25) gives us

$$\mathbb{E}f(\mathbf{x}^{k+1}) \leq (1 + 2a_1\gamma_k\delta_k)\mathbb{E}f(\mathbf{x}^k) - \gamma_k \mathbb{E}\|\nabla f(\mathbf{x}^k)\|_1 + \frac{Ld\gamma_k^2}{2} + 2a_2\gamma_k\delta_k. \tag{20}$$

Using Lemma 2, it follows that

$$\mathbb{E}f(\mathbf{x}^k) \leq \exp(\sum_{i=C}^{k} 2a_1\gamma_i\delta_i) \times \left(f(\mathbf{x}^C) + \sum_{i=C}^{k} (\frac{Ld}{2}\gamma_i^2 + 2a_2\gamma_i\delta_i) \right) \leq \exp(2a_1C_1) \left(f(\mathbf{x}^C) + \frac{LdC_2}{2} + 2a_2C_1 \right). \tag{21}$$

Once from (20) and using bound (21), we can get

$$\begin{aligned} \sum_{i=C}^{k} \gamma_i \|\nabla f(\mathbf{x}^i)\|_1 &\leq f(\mathbf{x}^C) - \mathbb{E}f(\mathbf{x}^{k+1}) + 2a_1 \sum_{i=C}^{k} \gamma_i\delta_i \mathbb{E}f(\mathbf{x}^i) + \sum_{i=C}^{k} \frac{Ld\gamma_i^2}{2} \\ &\leq f(\mathbf{x}^C) - \min f + \frac{LdC_2}{2} + 2a_1C_1 \cdot \exp(2a_1C_1) \cdot \left(f(\mathbf{x}^C) + \frac{LdC_2}{2} + 2a_2C_1 \right), \end{aligned}$$

where we used the fact $f(\mathbf{x}) \geq \bar{f}$ for any $\mathbf{x} \in \text{dom}(f)$. To complete the proof, we notice

$$(\sum_{i=1}^{k} \gamma_i) \times \mathbb{E}(\min_{1\leq i\leq k} \{\|\nabla f(\mathbf{x}^i)\|_1\}) \leq \sum_{i=1}^{k} \gamma_i \|\nabla f(\mathbf{x}^i)\|_1.$$

Proof of (11)**:** We turn to bounding $\mid \mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|_1 - \mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|_1 \mid$ as

$$\begin{aligned} &\mid \mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|_1 - \mathbb{E}\|\nabla f(\mathbf{x}^k)\|_1 \mid \\ &\quad = \mid \mathbb{E}(\|\nabla f(\mathbf{x}^{k+1})\|_1 - \|\nabla f(\mathbf{x}^k)\|_1) \mid \\ &\leq \mathbb{E}\|\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)\|_1 \leq \sqrt{d}L\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \\ &\quad \leq \sqrt{d}L\gamma_k\|\mathbf{g}^k\| \leq Ld\gamma_k. \end{aligned} \tag{22}$$

Using Lemma 3 with $\gamma_k \to h_k$, $Ld \to c$ and $\mathbb{E}\|\nabla f(\mathbf{x}^k)\|_1 \to \alpha_k$, we immediately obtain $\lim_k \mathbb{E}\|\nabla f(\mathbf{x}^k)\|_1 = 0$.

E Proof of Theorem 1

It is easy to see $\mathbf{v}^k = \mathbf{g}^k = \frac{\sum_{l=1}^{n_k} \nabla f_{i_{k,l}}(\mathbf{x}^k)}{n_k}$. With the Markov's inequality and Jensen's inequality, we are then led to

$$\begin{aligned} &\mathcal{P}\{[\text{Sign}(\mathbf{g}_j^k) \neq [\text{Sign}(\nabla f(\mathbf{x}^k))]_j\} \\ &\quad \leq \mathcal{P}\{\mid \mathbf{g}_j^k - [\nabla f(\mathbf{x}^k)]_j \mid \geq |[\nabla f(\mathbf{x}^k)]_j|\} \\ &\leq \frac{\mathbb{E}(|\mathbf{g}_j^k - [\nabla f(\mathbf{x}^k)]_j| \mid \chi^k)}{|[\nabla f(\mathbf{x}^k)]_j|} \\ &\quad \leq \frac{\sqrt{\mathbb{E}((\mathbf{g}_j^k - [\nabla f(\mathbf{x}^k)]_j)^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|} \\ &\leq \frac{\sqrt{\mathbb{E}(\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|} \\ &\quad \leq \frac{\sqrt{\mathbb{E}(\|f_{i_k}(\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\|^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j| \cdot \sqrt{n_k}} \\ &\leq \frac{\sqrt{\mathbb{E}(\|\nabla f_{i_k}(\mathbf{x}^k)\|^2 \mid \chi^k) - \|\nabla f(\mathbf{x}^k)\|^2}}{|[\nabla f(\mathbf{x}^k)]_j| \cdot \sqrt{n_k}}, \end{aligned} \tag{23}$$

where i_k is selected uniformly from $\{1, 2, \ldots, n\}$. With Lemma 1, we have

$$\begin{aligned} &\sqrt{\mathbb{E}(\|\nabla f_{i_k}(\mathbf{x}^k)\|^2 \mid \chi^k) - \|\nabla f(\mathbf{x}^k)\|^2} \leq \sqrt{\mathbb{E}(\|\nabla f_{i_k}(\mathbf{x}^k)\|^2 \mid \chi^k)} \leq \sqrt{\mathbb{E}(2\bar{L}(f_{i_k}(\mathbf{x}^k) - \min f_{i_k}) \mid \chi^k)} \\ &= \sqrt{2\bar{L}(f(\mathbf{x}^k) - \bar{f})} \leq \sqrt{\frac{\bar{L}}{2\bar{f}}} f(\mathbf{x}^k) + \sqrt{\frac{\bar{L}\bar{f}}{2}}, \end{aligned} \tag{24}$$

where we used the inequality $\sqrt{ab} \leq \frac{\mu}{2}a + \frac{b}{2\mu}$ with $a = f(\mathbf{x}^k) - \bar{f}$, $b = 2\bar{L}$ and $\mu = \sqrt{\frac{\bar{L}}{2\bar{f}}}$. Combining (23) and (24), we then derive

$$\sum_{i=1}^{d} |[\nabla f(\mathbf{x}^k)]_i| \cdot \mathcal{P}[\text{Sign}(\mathbf{v}_i^k) \neq \text{Sign}([\nabla f(\mathbf{x}^k)]_i)|\chi^k] \leq \sqrt{\frac{\bar{L}}{2\bar{f}}} \frac{1}{\sqrt{k}} f(\mathbf{x}^k) + \sqrt{\frac{\bar{L}\bar{f}}{2}} \frac{1}{\sqrt{k}} \tag{25}$$

Hence, here $C=1$. Using Lemma 4, to complete the proof, we notice

$$(\sum_{i=1}^{k}\gamma_i)\times\mathbb{E}(\min_{1\leq i\leq k}\{\|\nabla f(\mathbf{x}^i)\|_1\})\leq\sum_{i=1}^{k}\gamma_i\mathbb{E}\|\nabla f(\mathbf{x}^i)\|_1.$$

F Proof of Theorem 2

The PL property indicates

$$\|\nabla f(\mathbf{x}^k)\|_1\geq\frac{\|\nabla f(\mathbf{x}^k)\|}{\sqrt{d}}\geq\frac{2\nu}{\sqrt{d}}\sqrt{f(\mathbf{x}^k)-\min f}. \tag{26}$$

The convergence result $(f(\mathbf{x}^k))_{k\geq 0}$ is convergent a.s. still hold. Assume that $(\xi_k:=f(\mathbf{x}^k)-\min f)_{k\geq 0}$ converges to ξ a.s. Obviously, $\xi\geq 0$ a.s. and $(\sqrt{\xi_k})_{k\geq 0}$ converges to $\sqrt{\xi}$ a.s. Equation (26) tells us

$$\mathbb{E}\sqrt{\xi}=\liminf_k\mathbb{E}\sqrt{f(\mathbf{x}^k)-\min f}\leq\frac{\sqrt{d}}{2\nu}\liminf_k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|_1=0. \tag{27}$$

The nonnegativity of $\sqrt{\xi}$ means $\sqrt{\xi}=0$ a.s., which also indicates $\xi=0$ a.s.

Now we prove an important bound for the analysis. We claim

$$\delta:=\liminf_k\frac{\mathbb{E}\|\nabla f(\mathbf{x}^k)\|_1}{\mathbb{E}(f(\mathbf{x}^k)-\min f)}>0. \tag{28}$$

Otherwise, there exists $(k_j)_{j\geq 0}$ such that $\frac{\mathbb{E}\|\nabla f(\mathbf{x}^{k_j})\|_1}{\mathbb{E}(f(\mathbf{x}^{k_j})-\min f)}\to 0$. With (26), we then get $\frac{\mathbb{E}\sqrt{f(\mathbf{x}^{k_j})-\min f}}{\mathbb{E}(f(\mathbf{x}^{k_j})-\min f)}\to 0$. Using the Cauchy's inequality $\mathbb{E}\sqrt{f(\mathbf{x}^{k_j})-\min f}\leq\sqrt{\mathbb{E}(f(\mathbf{x}^{k_j})-\min f)}$, we then get $\frac{1}{\sqrt{\mathbb{E}(f(\mathbf{x}^{k_j})-\min f)}}\to 0$, which contradicts the fact $\xi=0$ a.s.

Back to (20) with (25) and (28), for k large enough, we are led to

$$\mathbb{E}\xi_{k+1}\leq(1+d\sqrt{\frac{2\bar{L}}{\bar{f}}}\frac{\gamma_k}{\sqrt{k}}-\delta\gamma_k)\mathbb{E}\xi_k+\frac{Ld\gamma_k^2}{2}+\sqrt{2\bar{L}\bar{f}}\frac{\gamma_k}{\sqrt{k}}+\min f\cdot d\sqrt{\frac{2\bar{L}}{\bar{f}}}\frac{\gamma_k}{\sqrt{k}}. \tag{29}$$

When k is large enough, $d\sqrt{\frac{2\bar{L}}{\bar{f}}}\frac{1}{\sqrt{k}}<\frac{\delta}{2}$ and we have

$$\mathbb{E}\xi_{k+1}\leq(1-\frac{D\delta}{2}\frac{1}{k})\mathbb{E}\xi_k+D_1\frac{1}{k^{\frac{3}{2}}}, \tag{30}$$

where $D_1:=\frac{Ld}{2}+\sqrt{2\bar{L}\bar{f}}+\min f\cdot d\sqrt{\frac{2\bar{L}}{\bar{f}}}$ and the fact $\gamma_k\leq\frac{1}{\sqrt{k}}$ is used.

Note $\exp(\sum_{i=1}^{k}-\frac{D\delta}{2}\frac{1}{k})\leq-\frac{\delta}{2}\ln(k+1)$. and

$$\sum_{i=1}^{k}\exp(-\sum_{j=i+1}^{k}\frac{1}{j})\frac{1}{i^{\frac{3}{2}}}\leq\sum_{i=1}^{k}\exp(-\sum_{j=i+1}^{k}\int_j^{j+1}\frac{1}{t}dt)\frac{1}{i^{\frac{3}{2}}}\leq\sum_{i=1}^{k}\frac{i+1}{k}\frac{1}{i^{\frac{3}{2}}}=O(\frac{1}{\sqrt{k}}).$$

We then prove the result with Lemma 2.

G Proof of Theorem 3

It is to see Sign $\left[\sum_{m=1}^{M}\text{Sign}(\mathbf{g}^{k,m})\right]=$ Sign $\left[\frac{\sum_{m=1}^{M}\text{Sign}(\mathbf{g}^{k,m})}{M}\right]$. Using Lemma 4, we can see $\mathbf{v}^k=\frac{\sum_{m=1}^{M}\text{Sign}(\mathbf{g}^{k,m})}{M}$ here. With direct computations, we have

$$\begin{aligned}&\mathcal{P}\{[\text{Sign}(\mathbf{v}_j^k)\neq[\text{Sign}(\nabla f(\mathbf{x}^k))]_j\}\\&\quad\leq\mathcal{P}\{|\ \mathbf{g}_j^k-[\nabla f(\mathbf{x}^k)]_j\ |\geq|[\nabla f(\mathbf{x}^k)]_j|\}=\mathcal{P}\{|\ \frac{\sum_{m=1}^{M}(\text{Sign}(\mathbf{g}^{k,m})-\nabla f(\mathbf{x}^k))_j}{M}\ |\geq|[\nabla f(\mathbf{x}^k)]_j|\}\end{aligned}$$

With Lemma 5, we have

$$
\begin{aligned}
&\mathcal{P}\{\mid \frac{\sum_{m=1}^{M}(\text{Sign}(\mathbf{g}^{k,m})-\nabla f(\mathbf{x}^k))_j}{M} \mid\geq |[\nabla f(\mathbf{x}^k)]_j|\}\\
&\leq \sum_{m=1}^{M}\mathcal{P}\{\mid (\text{Sign}(\mathbf{g}^{k,m})-\nabla f(\mathbf{x}^k))_j \mid\geq |[\nabla f(\mathbf{x}^k)]_j|\} = \sum_{m=1}^{M}\mathcal{P}\{\mid (\mathbf{g}^{k,m}-\nabla f(\mathbf{x}^k))_j \mid\geq |[\nabla f(\mathbf{x}^k)]_j|\}\\
&\leq M\frac{\mathbb{E}(|\mathbf{g}_j^{k,1}-[\nabla f(\mathbf{x}^k)]_j| \mid \chi^k)}{|[\nabla f(\mathbf{x}^k)]_j|} \leq M\frac{\sqrt{\mathbb{E}((\mathbf{g}_j^{k,1}-[\nabla f(\mathbf{x}^k)]_j)^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|}\\
&\leq M\frac{\sqrt{\mathbb{E}(\|\mathbf{g}^{k,1}-\nabla f(\mathbf{x}^k)\|^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|} \leq M\frac{\sqrt{\mathbb{E}(\|f_{i_k}(\mathbf{x}^k)-\nabla f(\mathbf{x}^k)\|^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|\cdot\sqrt{n_k}}\\
&\leq M\frac{\sqrt{\mathbb{E}(\|\nabla f_{i_k}(\mathbf{x}^k)\|^2 \mid \chi^k)-\|\nabla f(\mathbf{x}^k)\|^2}}{|[\nabla f(\mathbf{x}^k)]_j|\cdot\sqrt{n_k}},
\end{aligned}
$$

Hence, we are led to

$$
\sum_{i=1}^{d}|[\nabla f(\mathbf{x}^k)]_i|\cdot\mathcal{P}[\text{Sign}(\mathbf{v}_i^k)\neq\text{Sign}([\nabla f(\mathbf{x}^k)]_i)|\chi^k] \leq M\sqrt{\frac{\bar{L}}{2\bar{f}}}\frac{1}{\sqrt{k}}f(\mathbf{x}^k)+M\sqrt{\frac{\bar{L}\bar{f}}{2}}\frac{1}{\sqrt{k}} \tag{31}
$$

In this theorem, $C=1$.

H Proof of Proposition 3

With the result in Proof of Theorem 3, we also have

$$
\mathbb{E}\xi_{k+1}\leq(1-\frac{\delta}{2}\frac{1}{k^q})\mathbb{E}\xi_k+D\frac{1}{k^{q+\frac{1}{2}}} \tag{32}
$$

for k large enough and $\xi_k := f(\mathbf{x}^k)-\min f$ and $D>0$ is a constant. The following proofs are almost identical to proofs of SIGNSGD with PŁ.

I Proof of Theorem 4

It is easy to see $\mathbf{v}^k=\mathbf{g}^k=\frac{\sum_{l=1}^{n_k}\nabla f_{i_{k,l}}(\mathbf{x}^k)}{n_k}$. The Markov's inequality and Jensen's inequality together with (5) and (6) yield

$$
\begin{aligned}
&\mathcal{P}\{[\text{Sign}(\mathbf{g}_j^k)\neq[\text{Sign}(\nabla f(\mathbf{x}^k))]_j\}\leq\mathcal{P}\{\mid \mathbf{g}_j^k-[\nabla f(\mathbf{x}^k)]_j \mid\geq|[\nabla f(\mathbf{x}^k)]_j|\}\\
&\leq\frac{\mathbb{E}(|\mathbf{g}_j^k-[\nabla f(\mathbf{x}^k)]_j| \mid \chi^k)}{|[\nabla f(\mathbf{x}^k)]_j|}\leq\frac{\sqrt{\mathbb{E}((\mathbf{g}_j^k-[\nabla f(\mathbf{x}^k)]_j)^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|}\\
&\leq\frac{\sqrt{\mathbb{E}(\|\mathbf{g}^k-\nabla f(\mathbf{x}^k)\|^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|}\leq\frac{\sqrt{\mathbb{E}(\|\mathbf{g}^k-\mathbb{E}\mathbf{g}^k\|^2 \mid \chi^k)+\mathbb{E}(\|\mathbb{E}\mathbf{g}^k-\nabla f(\mathbf{x}^k)\|^2 \mid \chi^k)}}{|[\nabla f(\mathbf{x}^k)]_j|}\\
&\leq\frac{\sqrt{\mathbb{E}(\|\mathbf{g}^k-\mathbb{E}\mathbf{g}^k\|^2 \mid \chi^k)}+C_1(d)u_k}{|[\nabla f(\mathbf{x}^k)]_j|}\leq\frac{\sqrt{\mathbb{E}(\|\mathbf{h}(i_k;u_k)(\mathbf{x}^k)-\mathbb{E}\mathbf{h}(i_k;u_k)(\mathbf{x}^k)\|^2 \mid \chi^k)}/\sqrt{n_k}+C_1(d)u_k}{|[\nabla f(\mathbf{x}^k)]_j|}\\
&\leq\frac{\sqrt{\mathbb{E}(\|\mathbf{h}(i_k;u_k)(\mathbf{x}^k)\|^2 \mid \chi^k)}/\sqrt{n_k}+C_1(d)u_k}{|[\nabla f(\mathbf{x}^k)]_j|}\leq\frac{C_1(d)u_k+\sqrt{C_2(d)}\cdot u_k/\sqrt{n_k}+\sqrt{C_3(d)}\mathbb{E}(\|\nabla f_{i_k}(\mathbf{x})\| \mid \chi^k)/\sqrt{n_k}}{|[\nabla f(\mathbf{x}^k)]_j|}\\
&\leq\frac{C_1(d)u_k+\sqrt{C_2(d)}\cdot u_k/\sqrt{n_k}+\sqrt{C_3(d)}(\sqrt{\frac{\bar{L}}{2\bar{f}}}f(\mathbf{x}^k)+\sqrt{\frac{\bar{L}\bar{f}}{2}})/\sqrt{n_k}}{|[\nabla f(\mathbf{x}^k)]_j|}
\end{aligned} \tag{33}
$$

where i_k is selected uniformly from $\{1,2,\dots,n\}$. Using Lemma 4, we then proved the result.

J Proof of Proposition 5

When k is large enough, we have

$$
\mathbb{E}\xi_{k+1}\leq(1-\frac{D\delta}{2}\frac{1}{k})\mathbb{E}\xi_k+D_1\frac{1}{k^{\frac{3}{2}}}+D_2\frac{1}{k^{p+1}}, \tag{34}
$$

where δ is given by (28), and $\xi_k := f(\mathbf{x}^k) - \min f$, and $D_1, D_2 > 0$ are constants.

If $p \geq \frac{1}{2}$, (34) then gives

$$\mathbb{E}\xi_{k+1} \leq (1 - \frac{D\delta}{2}\frac{1}{k^q})\mathbb{E}\xi_k + (D_1 + D_2)\frac{1}{k^{\frac{3}{2}}}.$$

Like previous analysis, we then get the result.

If $0 < p < \frac{1}{2}$, (34) then gives

$$\mathbb{E}\xi_{k+1} \leq (1 - \frac{D\delta}{2}\frac{1}{k})\mathbb{E}\xi_k + (D_1 + D_2)\frac{1}{k^{p+1}}.$$

To get the second result, we observe

$$\sum_{i=1}^{k} \exp(-\sum_{j=i+1}^{k}\frac{1}{j})\frac{1}{i^{1+p}} \leq \sum_{i=1}^{k} \exp(-\sum_{j=i+1}^{k}\int_{j}^{j+1}\frac{1}{t}dt)\frac{1}{i^{1+p}} \leq \sum_{i=1}^{k}\frac{i+1}{k}\frac{1}{i^{1+p}} = O(\frac{1}{k^p}).$$

References

[Robbins and Siegmund, 1971] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.